Methods for Detecting Interactions Between Genetic Polymorphisms and Prenatal Environment Exposure with a Mother-Child Design

Shuang Wang,[1*] Tian Zheng,[2] Stephen Chanock,[3] Wieslaw Jedrychowski,[4,5] and Frederica P. Perera[4]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York*
[2]*Department of Statistics, Columbia University, New York, New York*
[3]*National Cancer Institute, Bethesda, Maryland*
[4]*Columbia Center for Children's Environmental Health, Mailman School of Public Health, Columbia University, New York, New York*
[5]*Department of Epidemiology and Preventive Medicine, College of Medicine, Jagiellonian University, Krakow, Poland*

Prenatal exposures such as polycyclic aromatic hydrocarbons and early postnatal environmental exposures are of particular concern because of the heightened susceptibility of the fetus and infant to diverse environmental pollutants. Marked inter-individual variation in response to the same level of exposure was observed in both mothers and their newborns, indicating that susceptibility might be due to genetic factors. With the mother-child pair design, existing methods developed for parent-child trio data or random sample data are either not applicable or not designed to optimally use the information. To take full advantage of this unique design, which provides partial information on genetic transmission and has both maternal and newborn outcome status collected, we developed a likelihood-based method that uses both the maternal and the newborn information together and jointly models gene-environment interactions on maternal and newborn outcomes. Through intensive simulation studies, the proposed method has demonstrated much improved power in detecting gene-environment interactions. The application on a real mother-child pair data from a study conducted in Krakow, Poland, suggested four significant gene-environment interactions after multiple comparisons adjustment. *Genet. Epidemiol.* 34 : 125–132, 2010.

Key words: gene-environment interaction; mother-child pair design; likelihood ratio test

Contract grant sponsor: National Institute of Environmental Health Sciences; Contract grant numbers: P01 009600; R01 ES08977; R01 ES111158; R01 ES012468; ES09089; Contract grant sponsor: US Environmental Protection Agency; Contract grant numbers: R827027; R8260901; Contract grant sponsor: Herbert Irving Cancer Center; Contract grant number: 5P30 CA 13696-23; Contract grant sponsors: New York Community Trust; Gladys & Roland Harriman Foundation; Blanchette Hooker Rockefeller Fund.

*Correspondence to: Shuang Wang, PhD, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, Room 630, New York, NY 10032. E-mail: sw2206@columbia.edu

Received 31 March 2009; Accepted 18 May 2009
Published online 6 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20441

INTRODUCTION

The impact of environmental exposures on childhood health is a crucial and understudied area of research. There has been growing evidence showing marked inter-individual variation in response to the same level of prenatal or postnatal environmental exposures in mothers and their newborns, which indicates susceptibility due to genetic factors. Complex childhood disorders including physical development problems, neurodevelopment disorders and childhood asthma are believed to result from interactions between multiple genetic factors and environmental factors. To better understand the causes of childhood disorders in children and the impact of early-life exposures from common urban pollutants on childhood disorders, the mother-child design has been applied [Wang et al., 2008]. Conventionally, association designs such as case-control designs, case-parents designs or case-only designs are utilized to detect gene-environment interactions [Schaid, 1999; Umbach and Weinberg, 2000; Khoury and Flanders, 1996; Albert et al., 2001]. There has been much methodological research conducted on detecting gene-environment interactions under these designs [Yang and Khoury, 1997; Thomas, 2004; Dempfle et al., 2008]. The population-based case-control design has gained its popularity due to its cost efficiency and feasibility for late-onset diseases but has been criticized for possibly inducing spurious association due to population stratification [Risch, 2000]. The case-parents design, as a family-based design, controls well against population stratification. The case-only design was proposed to avoid possible bias due to inappropriate definitions of "control" but relies heavily on the assumption that the gene and the environmental exposure are independent in the population [Khoury and Flanders, 1996; Albert et al., 2001]. However, with mother-child pair data when paternal information is completely missing, existing methods developed for case-control, case-parents and case-only designs are either not applicable or not designed to optimally use information from such studies. The current genetic analysis of the mother-child pair data has mainly used standard linear regression or logistic regression approaches that ignore the interrelationship between mothers and their newborns, which is not the most efficient use of information [Wang et al., 2008]. Most recently, researchers have worked on methodology development with case-mother/control-mother designs [Shi et al., 2008; Chen et al., 2009a,b]. However, no work has been done on prospective cohort designs with mother-child pairs. Therefore, to take full advantage of this unique design with mother-child pairs that provides partial information on genetic transmission and has both maternal and newborn outcome status collected, we developed a likelihood-based method that uses both the maternal and the newborn information together and jointly models gene-environment interactions on maternal and newborn outcomes. The proposed method is able to take into account the dependence between maternal and newborn genotypes as well as the dependence between their phenotypes. Through extensive simulation studies and an application to the Krakow mother-child pair data, we demonstrated that the proposed method outperformed a naïve method that ignored the interrelationship between mothers and newborns in detecting gene-environment interactions.

The development of this likelihood-based method for the mother-child pair design was motivated by two parallel ongoing studies being conducted in NYC and Krakow, Poland, with a mother-child pair design, in which pregnant women were recruited and monitored with personal air samples and their newborns were followed. In both studies, pregnant women were eligible if they were not currently smoking, registered at prenatal health-care clinics, had lived at the present address for at least a year before the initial interview, were ≥18 years of age, had no history of illicit drug use, pregnancy-related diabetes or hypertension and had a valid estimate of gestational age [Anderson et al., 2000; Perera et al., 2004]. During the second or third trimesters of pregnancy, the women carried a backpack containing a portable personal exposure air monitor during the day and kept it near the bed at night during a consecutive 48-hr period for polycyclic aromatic hydrocarbon (PAH) measurements, a widespread class of pollutants commonly found in air, food and drinking water [International Agency for Research on Cancer, 1983]. PAHs bind covalently to DNA to form PAH-DNA adducts, an indicator of DNA damage. Carcinogen-DNA adducts represent a critical step in the carcinogenic pathways and thus can be considered as an informative biomarker of cancer risk [Bulay and Wattenberg, 1971; Rice and Ward, 1982; Vesselinovitch et al., 1975]. We applied the proposed method to detect gene-environment interactions on PAH-DNA adduct detectable/non-detectable from the Krakow study. The binary outcome under investigation was the presence of detectable PAH-DNA adducts. Four significant gene-environment interactions were identified after adjusting for multiple comparisons. The proposed method is readily applicable to detect gene-environment interactions on other childhood disorder outcomes once the data collection is finished.

MATERIAL AND METHODS

Assume a bi-allelic marker with a high-risk allele A and a low-risk allele a, which have frequencies p and $1-p$, respectively. Here the high-risk allele is the minor allele (defined as the less common allele in the cohort). For a prospective design with sample units being mother-child pairs, we have both genetic information and outcome status collected for mothers and their children. PAHs inhaled by the pregnant women can be transferred to the fetus. Therefore, prenatal maternal monitoring provides an estimate of fetal exposure. The likelihood of observing a set of n independent mother-child pairs with prospective outcome status from a prospective cohort study is

$$\prod_{i=1}^{n} \Pr(G_{\mathrm{c}_i}, G_{\mathrm{m}_i}, E_i, D_{\mathrm{c}_i}, D_{\mathrm{m}_i}), \quad i = 1, \ldots, n,$$

where G_{m} and G_{c} stand for the maternal and newborn genotypes, E stands for the environmental exposure of both the mother and the newborn and D_{m} and D_{c} stand for the outcome status of the mother and the newborn, respectively. In deriving this likelihood, we assume that: (1) maternal genotypes are independent of their exposures to common urban pollutants; (2) the child's outcome status is determined by his/her genotype, prenatal environmental exposure and their interaction as well as the maternal outcome status and (3) the maternal outcome status is determined by the maternal genotype, environmental exposure and their interaction. Hardy-Weinberg equilibrium and random mating are also assumed. We could write the log-likelihood as

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \log(\,\Pr(G_{\mathrm{c}_i}, G_{\mathrm{m}_i}, E_i, D_{\mathrm{c}_i}, D_{\mathrm{m}_i}))\\ &= \sum_{i=1}^{n} \log(\Pr(D_{\mathrm{c}_i}|G_{\mathrm{c}_i}, E_i, D_{\mathrm{m}_i})\Pr(D_{\mathrm{m}_i}|G_{\mathrm{m}_i}, E_i)\\ &\quad \times \Pr(E_i)\Pr(G_{\mathrm{c}_i}, G_{\mathrm{m}_i}))\\ &= \sum_{i=1}^{n} \log\Bigg(\Pr(D_{\mathrm{c}_i}|G_{\mathrm{c}_i}, E_i, D_{\mathrm{m}_i})\Pr(D_{\mathrm{m}_i}|G_{\mathrm{m}_i}, E_i)\\ &\quad \times \Pr(E_i) \sum_{G_{\mathrm{p}_i} \in (G_{\mathrm{c}_i}, G_{\mathrm{m}_i})} (\Pr(G_{\mathrm{c}_i}|G_{\mathrm{m}_i}, G_{\mathrm{p}_i})\Pr(G_{\mathrm{m}_i}, G_{\mathrm{p}_i}))\Bigg)\end{aligned}$$

where $G_{\mathrm{p}} \in (G_{\mathrm{c}}, G_{\mathrm{m}})$ stands for paternal genotypes that are compatible to the maternal and child genotypes observed. Logistic model is applied to model maternal penetrance $\Pr(D_{\mathrm{m}} = 1 \mid G_{\mathrm{m}}, E)$ and child penetrance $\Pr(D_{\mathrm{c}} = 1 \mid G_{\mathrm{c}}, E, D_{\mathrm{m}})$ with gene-environment interactions:

$$\begin{aligned}&\mathrm{logit}(\Pr(D_{\mathrm{m}_i} = 1|G_{\mathrm{m}_i}, E_i))\\ &\quad = \beta_{0\mathrm{m}} + \beta_{G_{\mathrm{m}}} G_{\mathrm{m}_i} + \beta_{E_{\mathrm{m}}} E_i + \beta_{G_{\mathrm{m}}E} G_{\mathrm{m}_i} \times E_i,\\ &\mathrm{logit}(\Pr(D_{\mathrm{c}_i} = 1|G_{\mathrm{c}_i}, E_i, D_{\mathrm{m}_i}))\\ &\quad = \beta_{0\mathrm{c}} + \beta_{G_{\mathrm{c}}} G_{\mathrm{c}_i} + \beta_{E_{\mathrm{c}}} E_i + \beta_{G_{\mathrm{c}}E} G_{\mathrm{c}_i} \times E_i + \beta_{D_{\mathrm{m}}} D_{m_i},\end{aligned}$$

where $\beta_{0\mathrm{c}}$ and $\beta_{0\mathrm{m}}$ are intercepts, $\beta_{G_{\mathrm{m}}}$ and $\beta_{G_{\mathrm{c}}}$ are the regression coefficients representing the main effects of the genetic polymorphism on the maternal and newborn outcomes, respectively, $\beta_{D_{\mathrm{m}}}$ is the regression coefficient representing the main effect of the maternal outcome status on the newborn outcome status, $\beta_{E_{\mathrm{m}}}$ and $\beta_{E_{\mathrm{c}}}$ are the regression coefficients representing the main effects of the environmental exposure on the maternal and newborn outcomes, respectively, and $\beta_{G_{\mathrm{m}}E}$ and $\beta_{G_{\mathrm{c}}E}$ are the regression coefficients representing the gene-environment interaction effects on the maternal and newborn outcomes, respectively. Here E is the binary indicator for the environmental exposure for both the mother and the newborn; G_{m} and G_{c} are the maternal and newborn genotype codings, which depend on the genetic model studied. Under the dominant genetic model, G_m or $G_c = 1$ for genotypes AA and Aa, G_m or $G_c = 0$ for genotypes aa. Under the recessive genetic model, G_m or $G_c = 1$ for genotypes AA, G_m or $G_c = 0$ for genotypes Aa and aa. Under the additive model, G_m or $G_c = 2$ for genotypes AA, G_m or $G_c = 1$ for genotypes Aa and G_m or $G_c = 0$ for genotypes aa.

To test the null hypothesis of no interaction between genetic polymorphisms and environmental exposures on the outcomes, we test the null hypothesis $H_0 : \beta_{G_cE} = \beta_{G_mE} = 0$. The corresponding alternative hypothesis is H_1: at least one of $\beta_{G_cE}, \beta_{G_mE}$ is not 0. We use the likelihood ratio test with 2 degrees of freedom (dfs) to test the null hypothesis.

Note that the two intercepts β_{0c} and β_{0m} are not free parameters. Instead, they are determined by the population prevalences of the studied outcome in the maternal cohort, pD_m, and that in the newborn cohort, pD_c. This is comparable with the procedure used in our previous work on gene-gene interactions in the association and linkage studies [Wang and Zhao, 2003, 2007]. We can obtain β_{0c} and β_{0m} by solving the following two equations:

$$\begin{aligned}
&\text{Pop prevalence of female adults in a certain age range}\\
&= pD_m\\
&= \Pr(D_m = 1)\\
&= \sum_{G_m,E} \Pr(D_m = 1|G_m, E)\Pr(G_m, E),
\end{aligned}$$

$$\begin{aligned}
&\text{Pop prevalence of children in a certain age range}\\
&= pD_c\\
&= \Pr(D_c{=}1) = \sum_{G_c,E,D_m} \Pr(D_c = 1|G_c, E, D_m)\Pr(G_c, E, D_m)\\
&= \sum_{G_c,E,D_m} \left(\Pr(D_c = 1|G_c, E, D_m) \sum_{G_m} \Pr(G_m, G_c, E, D_m) \right)\\
&= \sum_{G_c,E,D_m} \Bigg(\Pr(D_c = 1|G_c, E, D_m)\\
&\quad \times \sum_{G_m} (\Pr(D_m|G_m, E,)\Pr(G_m, G_c)\Pr(E)) \Bigg).
\end{aligned}$$

The same logistic model is used to model maternal and newborn penetrances $\Pr(D_m = 1 \mid G_m, E)$ and $\Pr(D_c = 1 \mid G_c, E, D_m)$. Note that β_{0m} is determined first by the first equation and is then plugged into the second equation to solve for β_{0c}.

RESULTS

SIMULATION STUDIES: TYPE I ERROR AND POWER

In this section, using simulations under a variety of models, the performance of the proposed method is compared to that of a naïve method [Wang et al., 2008]. The naïve method ignores the interrelationship between mothers and their newborns but treats the maternal outcome status and the newborn outcome status as independent outcomes. The naïve method simply applies three separate logistic regression models, and therefore is termed the "3-logit naïve method" for the rest of this article:

$$\begin{aligned}
&\mathrm{logit}(\Pr(D_{m_i} = 1|G_{m_i}, E_i))\\
&\quad = \beta_{10} + \beta_{11}G_{m_i} + \beta_{12}E_i + \beta_{13}G_{m_i} \times E_i,\\
&\mathrm{logit}(\Pr(D_{c_i} = 1|G_{c_i}, E_i))\\
&\quad = \beta_{20} + \beta_{21}G_{c_i} + \beta_{22}E_i + \beta_{23}G_{c_i} \times E_i,\\
&\mathrm{logit}(\Pr(D_{c_i} = 1|G_{m_i}, E_i))\\
&\quad = \beta_{30} + \beta_{31}G_{m_i} + \beta_{32}E_i + \beta_{33}G_{m_i} \times E_i.
\end{aligned}$$

The first logistic regression models maternal outcomes over $G_m \times E$ interaction, the second models newborn outcomes over $G_c \times E$ interaction and the third models newborn outcomes over $G_m \times E$ interaction with the reasoning that maternal genotypes modulate effects of PAHs on fetus growth. Note that this reasoning is not directly used in the proposed method, but is indirectly incorporated by assuming maternal outcome influences child's outcome. If any one of the three gene-environment interactions is significant using the Wald test to test if the regression coefficient of the interaction term is 0 with 1 df at the Bonferroni corrected significance level (to adjust for the fact that three tests are conducted), we conclude that the genetic polymorphism significantly interacts with the environmental exposure on the outcome.

SIMULATION PARAMETERS

In the simulation studies to evaluate power and Type I error rate, the total sample size was fixed at $N = 500$ mother-child pairs. To mimic the NYC and Krakow studies on the outcome of PAH-DNA adduct detectable/non-detectable, we considered that population prevalences in mother and newborn cohorts are both 30 and 60% (the PAH-DNA adduct detectable rate was about 60% in both mother and newborn cohorts in the Krakow study), the probability of environmental exposure is $P_{\mathrm{PAH}} = 0.3$ and 0.5 and different minor allele frequencies (MAF) are $p = 0.1$, 0.2 and 0.3. Simulation scenarios when pD_m was not equal to pD_c were also considered. The main effects of the maternal and newborn genetic polymorphism on the maternal and newborn outcomes, the main effect of the environmental exposure on both maternal and newborn outcomes and the main effect of the maternal outcome status on the newborn outcome status were fixed at $\mathrm{OR}_{G_m} = 1.5$, $\mathrm{OR}_{G_c} = 1.5$, $\mathrm{OR}_E = 1.5$ and $\mathrm{OR}_{D_m} = 1.5$. Therefore, the corresponding regression coefficients in the logistic models for penetrances are all log(1.5). Different levels of the maternal and/or newborn gene-environment interaction effects on the maternal and newborn outcomes were considered ranging from low to high, $\mathrm{OR}_{G_c \times E} = 1.0, 1.5, 2.0, 3.0, 4.0$ and $\mathrm{OR}_{G_m \times E} = 1.0, 1.5, 2.0, 3.0, 4.0$. Scenarios where either only maternal genotype interacts with the environment or only newborn genotype interacts with the environment were also considered.

SIMULATION SETUP

Each simulated study included N mother-child pairs. We simulated N maternal genotypes and N paternal genotypes based on the population allele frequencies and the assumptions of Hardy-Weinberg equilibrium and random mating. Newborn genotypes were generated based on Mendelian transmission and generated parental genotypes. However, paternal genotypes were discarded. Environmental exposures of mothers and newborns were generated based on a binomial distribution with prespecified proportion of exposure. Under selected parameter settings, for a mother-child pair, with simulated genotypes and environmental exposure, their outcome statuses were generated based on the maternal and newborn penetrances $\Pr(D_m = 1 \mid G_m, E)$ and $\Pr(D_c = 1 \mid G_c, E, D_m)$. The tests of interest (the proposed method and the 3-logit naïve method) were performed using the simulated data and the procedures were repeated 10,000 times to evaluate the Type I error rates and 1,000 times to evaluate powers.

TYPE I ERROR

In order to evaluate the Type I error rate for the proposed test, simulation was used to generate data under the null hypothesis of no interaction between genetic polymorphisms and environmental exposure, $H_0 : \beta_{G_cE} = \beta_{G_mE} = 0$, i.e., $H_0 : \mathrm{OR}_{G_cE} = \mathrm{OR}_{G_mE} = 1$. The simulation procedure was repeated 10,000 times. Type I error rates of the proposed method and the 3-logit naïve method were then estimated by the proportions of times that the null hypothesis of no interaction between genetic polymorphisms and the environmental exposure was rejected by these two methods. The Bonferroni correction was applied in the 3-logit naïve method to declare significance. Table I displays the Type I error rates to detect gene-environment interactions with the proposed method and the 3-logit naïve method under the dominant genetic model when different MAFs and population prevalences were assumed. Table III displays the results under the additive genetic model. In both cases, the nominal Type I error rate of 0.05 was well controlled for both the proposed method and the 3-logit naïve method. The closeness of the estimated values of 0.05 indicates a better performance.

POWER

Table II displays powers to detect gene-environment interactions with the proposed method and the 3-logit naïve method under the dominant genetic model and simulation parameters specified previously. The same power results under the scenarios where the effects of gene-environment interactions on maternal and newborn outcomes were the same were also plotted in Figure 1. Power was assessed with 1,000 simulations. It is clear that the proposed method consistently shows higher powers on all the scenarios considered. Especially when the MAF is low and the maternal and newborn population prevalences were high. When pD_m and pD_c were both set at 60% to mimic the PAH-DNA adduct detectable rate in Krakow, Poland, and when the MAF was set at 0.1, the 3-logit naïve method barely has any power even with the large effect size of the gene-environment interaction on both maternal and newborn outcomes, while the proposed method can achieve 70% power when the effect size is large. We observed a power increase of more than three-fold with the proposed method when the population prevalence was 60% and the genetic variant was rare with MAF equal to 0.1. When the genetic variant is common and the population prevalence is 60%, the proposed method has a power gain ranging from a little less than 20% when the interaction effect was large to about 50% when the interaction effect was modest. Similarly, when the maternal and newborn population prevalences were both set at 30%, the proposed method has a power gain ranging from less than 5% when the interaction effect was big to almost 40% when the interaction effect was modest across all three levels of MAF considered (Table II). That is, the power gain of the proposed method is greater when the effect size of gene-environment interaction is relatively modest. For the scenarios where there is gene-environment interaction on either maternal outcome or newborn outcome but not both, the proposed method also has consistently higher power than the 3-logit naïve method. Moreover, we notice that the power for both methods increases as MAF increases from rare to common as expected, yet the proposed method enjoys a greater gain in power.

Similar patterns were observed when assuming an additive genetic model and different levels of the maternal and newborn population prevalences. As displayed in Table III, the proposed method consistently demonstrates higher power than the 3-logit naïve method when PAH exposure was 30%. Similar power increases were also observed when PAH exposure was 50% (data not shown).

TABLE I. Type I error rates to detect gene-environment interactions at the 0.05 significance level for the proposed method and the 3-logit naïve method under the dominant genetic model when MAF was set at 0.1, 0.2 and 0.3, population prevalence was set at $pD_m = pD_c = 0.3$ and $pD_m = pD_c = 0.6$ and environment exposure was set at 30%

Pop. prev.	MAF[a]	Proposed method	3-Logit naïve method
$pD_m = 0.3$, $pD_c = 0.3$	0.1	0.046	0.043
	0.2	0.042	0.043
	0.3	0.042	0.042
$pD_m = 0.6$, $pD_c = 0.6$	0.1	0.046	0.043
	0.2	0.051	0.043
	0.3	0.046	0.042

The total sample size was fixed at $N = 500$ mother-child pairs. The simulation procedure was repeated 10,000 times.

[a] Minor allele frequency.

REAL DATA APPLICATION

The proposed method was applied to detect gene-environment interactions on PAH-DNA adduct detectable/non-detectable using the mother-child pair data from the Krakow study. Seventeen common genetic polymorphisms in candidate genes that play important roles in the metabolic activation of PAHs and PAH detoxification were selected (Table IV). The data set was previously analyzed on the continuous PAH-DNA adduct levels using a similar naïve method with three linear regression models [Wang et al., 2008]. Some significant gene-environment interactions were observed but none remains significant at the 0.05 significance level after multiple comparisons adjustment. Also note that the method of Alexandrov et al. [1992], which uses the HPLC-fluorescence method to detect B[*a*]P-DNA adducts (a proxy for PAH-DNA adducts) [Lederman et al., 2004] in maternal blood collected within 1 day postpartum and

TABLE II. Power to detect gene-environment interactions for the proposed method and the 3-logit naïve method under the dominant genetic model when MAF was set at 0.1, 0.2 and 0.3, population prevalences were set at $pD_m = pD_c = 0.3$ and $pD_m = pD_c = 0.6$, environmental exposure was set at 30% and the effects of gene-environment interactions on maternal and newborn outcomes ranged from odd ratios of 1.0 to 4.0

		MAF					
		0.1		0.2		0.3	
Pop. prev.	Gene-environment interaction effect size	Proposed	3-Logit	Proposed	3-Logit	Proposed	3-Logit
$pD_m = 0.3$, $pD_c = 0.3$	$\mathrm{OR}_{G_cE} = 4.0$, $\mathrm{OR}_{G_mE} = 4.0$	0.896	0.789	0.977	0.942	0.963	0.956
	$\mathrm{OR}_{G_cE} = 3.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.724	0.592	0.879	0.810	0.857	0.842
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.580	0.460	0.759	0.679	0.741	0.701
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 2.0$	0.371	0.282	0.507	0.421	0.507	0.444
	$\mathrm{OR}_{G_cE} = 1.5$, $\mathrm{OR}_{G_mE} = 1.5$	0.152	0.128	0.243	0.196	0.213	0.175
	$\mathrm{OR}_{G_cE} = 3.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.389	0.331	0.540	0.515	0.574	0.593
	$\mathrm{OR}_{G_cE} = 1.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.462	0.389	0.599	0.580	0.604	0.596
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.185	0.157	0.245	0.218	0.276	0.241
	$\mathrm{OR}_{G_cE} = 1.0$, $\mathrm{OR}_{G_mE} = 2.0$	0.223	0.184	0.310	0.261	0.297	0.269
$pD_m = 0.6$, $pD_c = 0.6$	$\mathrm{OR}_{G_cE} = 4.0$, $\mathrm{OR}_{G_mE} = 4.0$	0.694	0.193	0.905	0.751	0.948	0.898
	$\mathrm{OR}_{G_cE} = 3.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.536	0.144	0.766	0.572	0.820	0.750
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.429	0.120	0.615	0.461	0.702	0.601
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 2.0$	0.292	0.076	0.419	0.279	0.481	0.371
	$\mathrm{OR}_{G_cE} = 1.5$, $\mathrm{OR}_{G_mE} = 1.5$	0.145	0.045	0.201	0.123	0.225	0.170
	$\mathrm{OR}_{G_cE} = 3.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.259	0.068	0.416	0.308	0.533	0.468
	$\mathrm{OR}_{G_cE} = 1.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.352	0.103	0.494	0.389	0.525	0.504
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.149	0.048	0.215	0.143	0.257	0.216
	$\mathrm{OR}_{G_cE} = 1.0$, $\mathrm{OR}_{G_mE} = 2.0$	0.195	0.057	0.260	0.189	0.269	0.222

The total sample size was fixed at $N = 500$ mother-child pairs. The simulation procedure was repeated 1,000 times. MAF, minor allele frequency.

TABLE III. Power and Type I error rate to detect gene-environment interaction for the proposed method and the 3-logit naïve method under the additive genetic model when MAF was set at 0.1, 0.2 and 0.3, population prevalences were set at $pD_m = 0.3$ and $pD_c = 0.2$, environmental exposure was set at 30% and the effects of gene-environment interactions on maternal and newborn outcomes were ranging from odds ratio of 1.0 (for Type I error) to 4.0

		MAF					
		0.1		0.2		0.3	
Pop. prev.	Gene-environment interaction effect size	Proposed	3-Logit	Proposed	3-Logit	Proposed	3-Logit
$pD_m = 0.3$, $pD_c = 0.2$	$\mathrm{OR}_{G_cE} = 4.0$, $\mathrm{OR}_{G_mE} = 4.0$	0.965	0.888	0.992	0.990	1.0	1.0
	$\mathrm{OR}_{G_cE} = 2.0$, $\mathrm{OR}_{G_mE} = 3.0$	0.706	0.564	0.897	0.824	0.957	0.913
	$\mathrm{OR}_{G_cE} = 3.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.635	0.549	0.859	0.825	0.911	0.900
	$\mathrm{OR}_{G_cE} = 1.5$, $\mathrm{OR}_{G_mE} = 1.5$	0.223	0.167	0.345	0.275	0.349	0.315
	$\mathrm{OR}_{G_cE} = 1.0$, $\mathrm{OR}_{G_mE} = 1.0$	0.047	0.042	0.041	0.041	0.041	0.044

The total sample size was fixed at $N = 500$ mother-child pairs. Power was assessed based on 1,000 simulations. Type I error rate was assessed based on 10,000 simulations. MAF, minor allele frequency.

umbilical cord blood collected at delivery [Perera et al., 2004], has a coefficient of variation of 12% and a lower limit of detection of 0.25 adducts per 10^8 nucleotides. Samples below the limit of detection were assigned a value midway between the limit of detection and zero (0.125 adducts per 10^8 nucleotides) in previous analyses. We reanalyzed the data with the proposed method on the binary outcome PAH-DNA adduct detectable/non-detectable.

The data set consists of 307 mother-child pairs that have complete phenotype and environmental exposure information. Different cut points for the PAH summary measures were applied to obtain a binary PAH exposure, defined as PAH high or PAH low, in order to optimize the results. The PAH-DNA adduct detectable rate in both mothers and newborns in the Polish data is around 60%. To adjust for multiple comparisons, the q-value procedure based on the false discovery rate (FDR) was applied

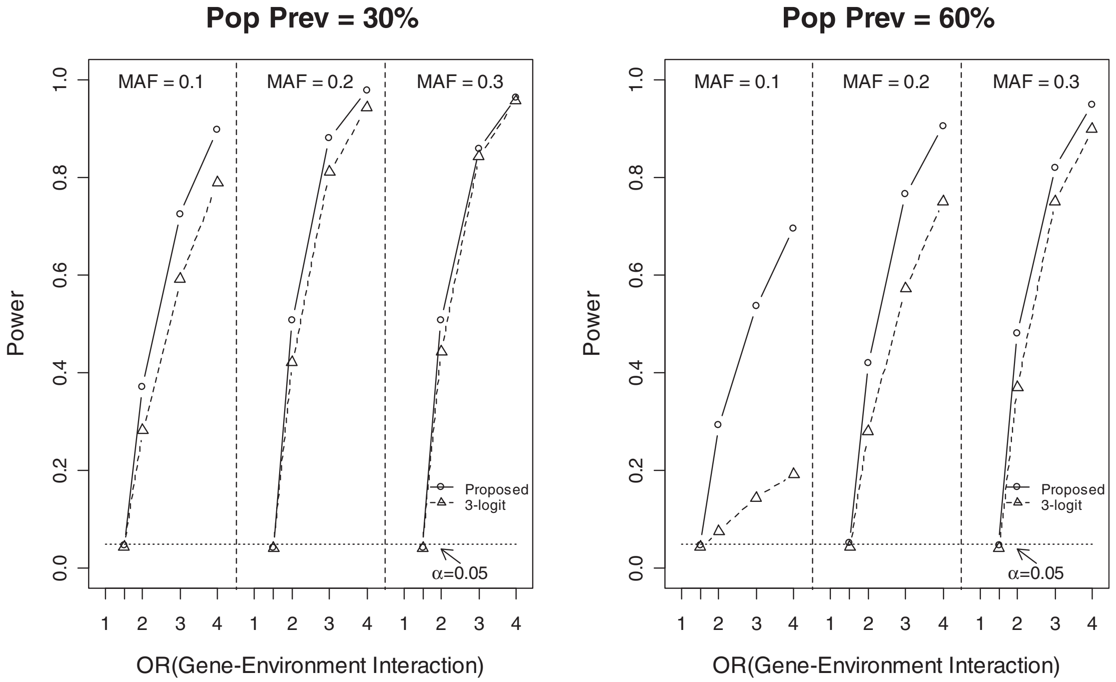

Fig. 1. Power to detect gene-environment interactions for the proposed method and the 3-logit naïve method under the dominant genetic model when MAF was set at 0.1, 0.2 and 0.3, population prevalence was set at $pD_m = pD_c = 30\%$ and $pD_m = pD_c = 60\%$, environmental exposure was set at 30% and the effects of gene-environment interactions on maternal and newborn outcomes were set as the same ranging from odds ratio of 1.0 (for Type I error) to 4.0. MAF, minor allele frequency.

TABLE IV. Chromosomal positions and gene locations of the 17 markers from the selected candidate genes

Gene	SNP	SNP rs num.	Alleles	Chr.	Position (bp)
CYP1A1	CYP1A1-78	rs2198843	C/G	15	72,788,283
	CYP1A1-109	rs1456432	A/G	15	72,790,104
	CYP1A1-06	rs4646903	C/T	15	72,798,694
	CYP1A1-15	rs4646421	T/C	15	72,803,245
	CYP1A1-14	rs2606345	T/G	15	72,804,229
	CYP1A1-83	rs7495708	G/A	15	72,806,896
	CYP1A1-81	rs2472299	C/T	15	72,820,453
CYP1B1	CYP1B1-66	rs162549	T/A	2	38,148,960
	CYP1B1-06	rs1056837	T/C	2	38,151,654
	CYP1B1-05	rs1056836	G/C	2	38,151,707
	CYP1B1-74	rs162560	A/G	2	38,153,019
	CYP1B1-04	rs10012	C/G	2	38,155,894
	CYP1B1-03	rs2617266	C/T	2	38,156,048
GSTT2	GSTT2-02	rs2719	T/G	22	22,629,757
	GSTT2-01	rs1622002	G/A	22	22,630,580
	GSTT2-03	rs140194	G/A	22	22,655,095
GSTM1	GSTM1-02	SNP500Cancer ID: GSTM1-02	Gene deletion	22	22,629,757

SNP, single nucleotide polymorphism.

[Storey and Tibshirani, 2003]. Based on our understanding of the biological basis of PAH-DNA adduct detectable/non-detectable, we considered that a dominant model might be the most appropriate.

Table V presents the q-values (those with q-values ≤ 0.05 were displayed in bold font) and estimates of the gene-environment interactions of 17 markers using both the proposed method and the 3-logit naïve method. For each marker, the most significant q-value among the three from the 3-logit models was displayed together with the regression coefficient estimate of the gene-environment interaction term in the corresponding model. We observed the same pattern in q-values using the two methods, but the proposed method outperformed the 3-logit naïve

TABLE V. Estimates of the interaction effects and *q*-value results at 17 candidate markers with the proposed method and the 3-logit naïve method with the PAH high/low cut point at 40%

	Proposed			3-Logit[a]	
Marker	$\hat{\beta}_{G_mE}$	$\hat{\beta}_{G_cE}$	*q*-Value	$\hat{\beta}_{\text{inter}}$	*q*-Value
CYP1A1-78	1.327	0.205	0.104	2.100	0.238[b]
CYP1A1-109	1.589	1.122	**0.034**	1.623	0.200[c]
CYP1A1-06	0.795	1.418	**0.034**	2.185	0.101[b]
CYP1A1-15	1.274	0.053	0.104	1.363	0.238[c]
CYP1A1-14	2.026	0.965	**0.034**	2.083	0.200[c]
CYP1A1-83	0.380	−0.980	0.107	−0.890	0.395[d]
CYP1A1-81	1.423	1.896	0.070	1.455	0.238[d]
CYP1B1-66	0.083	−0.959	0.107	−1.699	0.238[d]
CYP1B1-06	0.126	−0.626	0.149	−1.056	0.334[d]
CYP1B1-05	−0.155	0.565	0.106	1.608	0.238[d]
CYP1B1-74	−0.583	0.698	0.104	1.655	0.238[d]
CYP1B1-04	−0.716	0.575	0.104	1.389	0.238[d]
CYP1B1-03	−0.574	0.838	0.106	1.552	0.238[d]
GSTT2-02	4.946	−1.004	0.104	−0.268	0.598[d]
GSTT2-01	1.023	−0.322	0.104	1.154	0.238[c]
GSTT2-03	2.050	1.250	**0.008**	2.089	0.101[c]
GSTM1-02	−0.639	−0.031	0.149	−1.058	0.238[b]

Bold font indicates *q*-values$\leq$ 0.05. PAH, polycyclic aromatic hydrocarbon.

[a] For the 3-logit naïve method, the most significant *q*-value among the *q*-values from the three logit models was presented for each marker together with the regression coefficient estimate of the gene-environment interaction term in the corresponding model.

[b] Maternal marker∗environment interaction on newborn outcome.

[c] Maternal marker∗environment interaction on maternal outcome.

[d] Newborn marker∗environment interaction on newborn outcome.

method in general for this data set. Note that the 40% PAH cut point gave the most optimal results (that is, 40% PAH high vs. 60% PAH low) although all the PAH cut points tried suggested the better performance of the proposed method. Therefore, Table V displays results with environmental exposure defined as 40% PAH high vs. 60% PAH low. Four markers, *CYP1A1-109*, *CYP1A1-06*, *CYP1A1-14* and *GSTT2-03*, significantly interact with the environmental exposure at the 0.05 FDR level after multiple comparisons adjustment with the proposed method while no significance was observed at the 0.05 FDR level after multiple comparisons adjustment with the 3-logit naïve method. The same marker, *CYP1A1-14*, was previously observed to interact with the environmental exposure on the continuous PAH-DAN adduct measures at the 0.05 significance level before multiple comparisons adjustment, but did not remain significant after multiple comparisons adjustment. We observed the same pattern in the estimates of the gene-environment interactions using the two methods as well. For example, at the marker *CYP1A1-109* that remained significant after multiple comparisons adjustment, one of the three models from the 3-logit naïve method that models maternal outcome over $G_m \times E$ interaction showed the most significant result among the three, while the proposed method also had bigger estimated effect of maternal genotype by environment interaction on the maternal outcomes than the estimated effect of newborn genotype by environment interaction on the newborn outcome, i.e. $\hat{\beta}_{G_mE} > \hat{\beta}_{G_cE}$. Similar patterns were observed on the other three markers that significantly interact with the environmental exposure at the 0.05 FDR level.

The proposed method is readily applicable to other outcomes such as mental development problems or asthma from both the NYC study and the Krakow study once the data collection is finished. We expect the proposed method to have even greater power than the 3-logit naïve method for the complex outcomes mentioned above as the effect sizes for the complex outcomes are usually small to modest.

DISCUSSION

In this study, we proposed a likelihood-based method to detect gene-environment interactions with the mother-child pair design. The development of this method was motivated by the two parallel ongoing studies being conducted in NYC and Krakow, Poland, where the purpose of the studies is to understand the impact of environmental exposures on childhood health. Pregnant women were recruited and their newborns were followed up. Marked inter-individual variation in response to the same level of exposure was observed, indicating that susceptibility might be due to genetic factors, i.e. the existence of gene-environment interactions. Therefore, we focused our study on modeling gene-environment interactions. As existing methods are either not applicable or not designed to optimally use the information from such mother-child pair designs, we developed a likelihood-based method that uses both the maternal and the newborn information together and jointly models gene-environment interactions on maternal and newborn outcomes. Under this likelihood framework, one can also model data combined from parent-child trios and mother-child pairs using the proposed method. The proposed method imposes an underlying assumption that gene and environment are independent. Although we believe this assumption to be appropriate for the proposed model, which is for prospective cohort studies, we need to be cautious with such an assumption for case-control studies. In this prospective cohort study, eligible pregnant women living in the targeted geographic areas were recruited. If this is a case-control study and subjects were recruited based on some disease status, and the disease status is related to the genetics and environment, the assumption of gene and environment being independent might be violated, which can lead to biased parameter estimates of gene-environment interactions [Mukherjee and Chatterjee, 2008].

The simulation results illustrated the feasibility and power of the proposed method. The proposed method that jointly models gene-environment interactions on maternal and newborn outcomes has higher power to detect gene-environment interactions than the 3-logit naïve method that models gene-environment interactions on maternal and newborn outcomes separately. In the application to the real data on PAH-DNA adduct detectable/non-detectable from the Krakow study, although similar patterns were observed using both the proposed and the 3-logit naïve methods, the proposed method suggested four significant interactions at the 0.05 FDR level after adjusting for multiple comparisons while the 3-logit naïve method suggested no significant interaction after adjusting for multiple comparisons at the 0.05 FDR level. As data on early postnatal exposures are also being collected in both the NYC and the Krakow studies, the proposed method (here used prenatal exposures involving maternal-fetal transfer) can be easily adopted to model the postnatal exposure where different environment exposure measures for mother and newborn will be entered in the model.

We concentrated only on the mother-child pairs with complete phenotype and environmental exposure information. Note that there is missingness in phenotypes, prenatal environmental exposure measures and genotypes. Methods that could impute missing genotypes and phenotypes may be applied [Scheet and Stephens, 2006; Marchini et al., 2007; Browning and Browning, 2007]. In addition, measurement errors may exist in PAH measures and PAH-DNA adduct measures. An extension of the currently proposed method to detect gene-environment interactions with measurement error incorporated and with the missing data problem considered will be our future study. Moreover, we did not consider dependence among markers but treated them as independent markers, which has been previously found in simulation studies to produce reasonable results and may lead to conservative estimation of the FDR [Storey and Tibshirani, 2003; Fernando et al., 2004]. We have implemented the proposed method in R, and the R program is available upon request from the first author. We are preparing to extend the current method, which focuses only on the interactions between a single marker and an environmental exposure, to a method that models the interactions between environment and multiple markers, especially multiple markers from one candidate gene.

ACKNOWLEDGMENTS

Support by the National Institute of Environmental Health Sciences (grants P01 009600, R01 ES08977, R01 ES111158, R01 ES012468 and ES09089), the US Environmental Protection Agency (grants R827027 and R8260901), the Herbert Irving Cancer Center, Core grant 5P30 CA 13696-23, the New York Community Trust, the Gladys & Roland Harriman Foundation and Trustees of the Blanchette Hooker Rockefeller Fund is gratefully acknowledged. All authors declare no conflict of interest on this work.

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693.

Alexandrov K, Rojas M, Geneste O, Castegnaro M, Camus AM, Petruzzelli S, Giuntini C, Bartsch H. 1992. An improved fluorometric assay for dosimetry of benzo(*a*)pyrene diol-epoxide-DNA adducts in smokers' lung: comparisons with total bulky adducts and aryl hydrocarbon hydroxylase activity. Cancer Res 52:6248–6253.

Anderson LM, Diwan BA, Fear NT, Roman E. 2000. Critical windows of exposure for children's health: cancer in human epidemiological studies and neoplasms in experimental animal models. Environ Health Perspect 108:573–594.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097.

Bulay OM, Wattenberg LW. 1971. Carcinogenic effects of polycyclic hydrocarbocarcinogens administrated to mice during pregnancy on the progeny. J Natl Cancer Inst 46:397–402.

Chen JB, Zheng HT, Wilson ML. 2009a. Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications. Genet Epidemiol, in press.

Chen JB, Zheng HT, Wilson ML, Kraft P. 2009b. Testing Hardy-Weinberg equilibrium using mother-child case-control samples. Genet Epidemiol, in press.

Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schafer H. 2008. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16:1164–1172.

Fernando RL, Nettleton D, Southey BR, Dekkers JC, Rothschild MF, Soller M. 2004. Controlling the proportion of false positive in multiple dependent tests. Genetics 166:611–619.

International Agency for Research on Cancer. 1983. Polynuclear aromatic compounds. Part 1. Chemical, environmental, and experimental data. IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Lyon, France: International Agency for Research on Cancer. p 1–453.

Khoury MJ, Flanders WD. 1996. Nontraditional epidemiologic approach in the analysis of gene environment interaction: case-control studies with no controls! Am J Epidemiol 144:207–213.

Lederman SA, Rauh V, Weiss L, Stein JL, Hoepner LA, Perera FP. 2004. The effects of the World Trade Center event on birth outcomes among term deliveries at three lower Manhattan hospitals. Environ Health Perspect 112:1772–1778.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694.

Perera FP, Tang D, Tu YT, Cruz LA, Borjas M, Bernert T, Whyatt RM. 2004. Biomarkers in maternal and newborn blood indicate heightened fetal susceptibility to procarcinogenic DNA damage. Environ Health Perspect 112:1133–1136.

Rice JM, Ward JM. 1982. Age dependence of susceptibility to carcinogenesis in the nervous system. Ann N Y Acad Sci 381:274–289.

Risch N. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Schaid DJ. 1999. Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644.

Shi M, Umbach DM, Vermeulen SH, Weinberg CR. 2008. Making the most of case-mother/control-mother studies. Am J Epidemiol 168:541–547.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide studies. Proc Natl Acad Sci 100:9440–9445.

Thomas DC. 2004. Statistical Methods in Genetic Epidemiology. New York: Oxford University Press. p 320–326.

Umbach DM, Weinberg CR. 2000. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66:251–261.

Vesselinovitch SC, Kyriazis AP, Mihailovich N, Rao KV. 1975. Conditions modifying development of tumors in mice at various sites by benzo(*a*)pyrene. Cancer Res 35:2948–2953.

Wang S, Zhao HY. 2003. Sample size needed to detect gene-gene interactions using association designs. Am J Epidemiol 158:899–914.

Wang S, Zhao HY. 2007. Sample size needed to detect gene-gene interactions using linkage analysis. Ann Hum Genet 71:828–842.

Wang S, Chanock S, Tang D, Li ZG, Jedrychowski W, Perera FP. 2008. An assessment of interactions between PAH exposure and genetic polymorphisms on PAH-DNA adducts in African American, Dominican, and Caucasian mothers and newborns. Cancer Epidemiol Biomarkers Prev 17:405–413.

Yang Q, Khoury MJ. 1997. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 19:33–43.